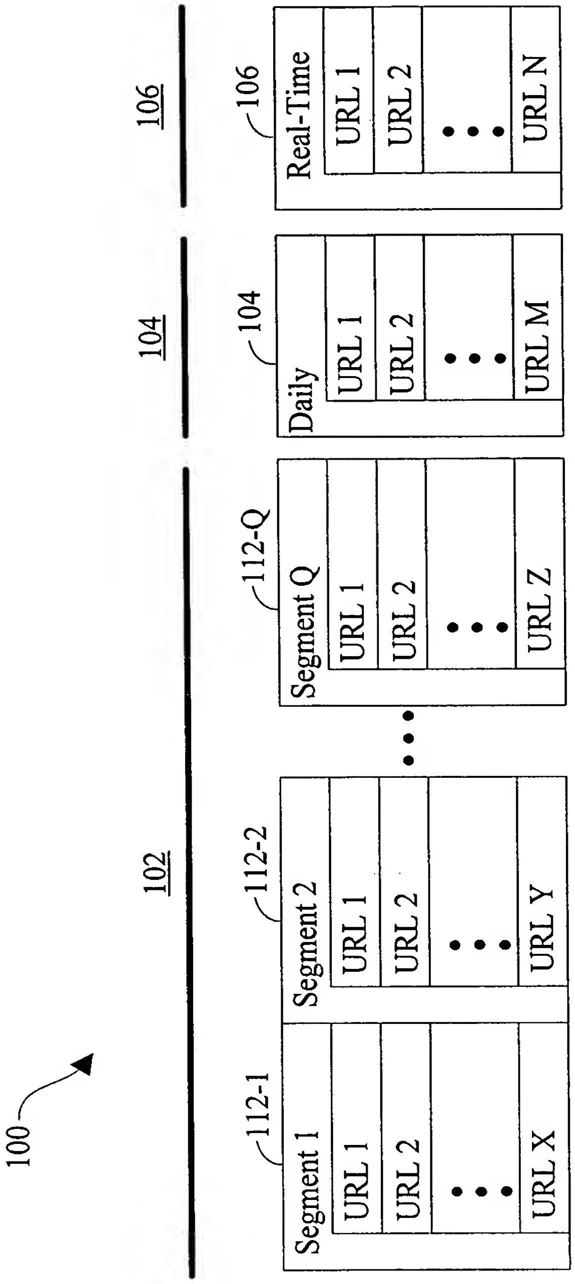

100
102
104
106
112-1
Segment 1
URL 1
URL 2
URL X
112-2
Segment 2
URL 1
URL 2
URL Y
112-Q
Segment Q
URL 1
URL 2
URL Z
104
Daily
URL 1
URL 2
URL M
106
Real-Time
URL 1
URL 2
URL N

Fig. 1

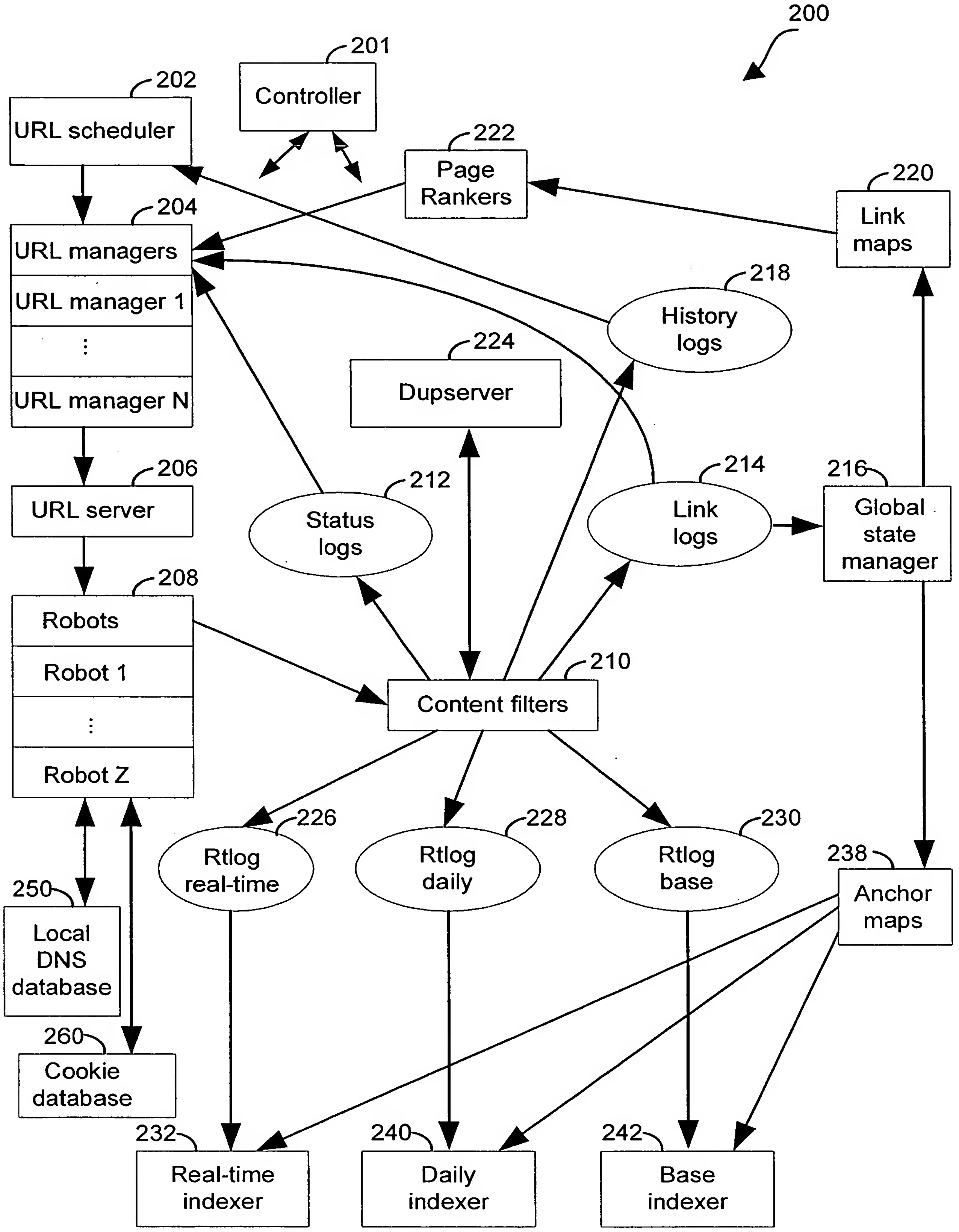
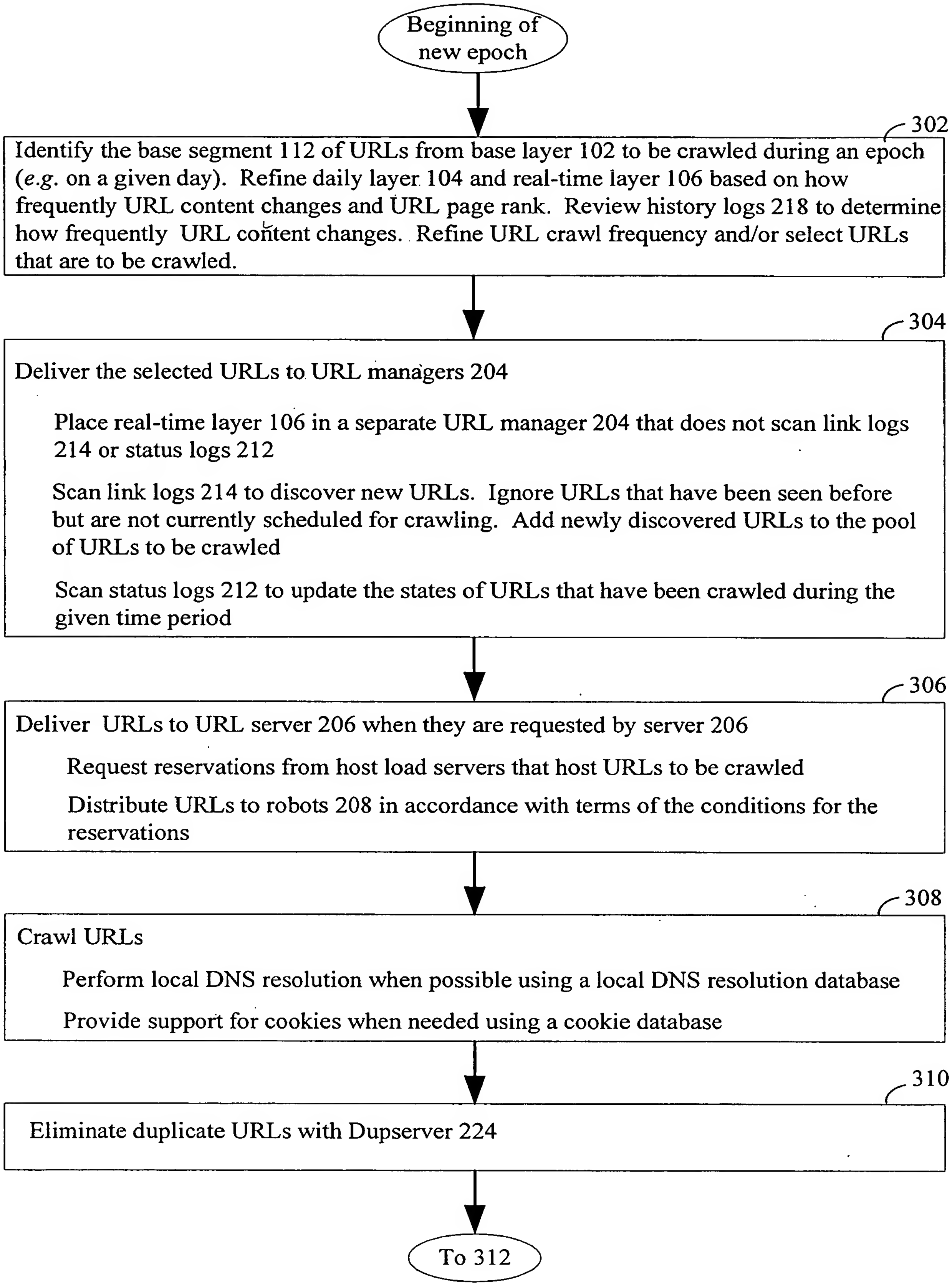

200
201
Controller
202
URL scheduler
222
Page Rankers
220
Link maps
204
URL managers
URL manager 1
⋮
URL manager N
218
History logs
224
Dupserver
206
URL server
212
Status logs
214
Link logs
216
Global state manager
208
Robots
Robot 1
⋮
Robot Z
210
Content filters
226
Rtlog real-time
228
Rtlog daily
230
Rtlog base
238
Anchor maps
250
Local DNS database
260
Cookie database
232
Real-time indexer
240
Daily indexer
242
Base indexer

Fig. 2



Identify the base segment 112 of URLs from base layer 102 to be crawled during an epoch (*e.g.* on a given day). Refine daily layer 104 and real-time layer 106 based on how frequently URL content changes and URL page rank. Review history logs 218 to determine how frequently URL content changes. Refine URL crawl frequency and/or select URLs that are to be crawled.

Deliver the selected URLs to URL managers 204

Place real-time layer 106 in a separate URL manager 204 that does not scan link logs 214 or status logs 212

Scan link logs 214 to discover new URLs. Ignore URLs that have been seen before but are not currently scheduled for crawling. Add newly discovered URLs to the pool of URLs to be crawled

Scan status logs 212 to update the states of URLs that have been crawled during the given time period

Deliver URLs to URL server 206 when they are requested by server 206

Request reservations from host load servers that host URLs to be crawled

Distribute URLs to robots 208 in accordance with terms of the conditions for the reservations

Crawl URLs

Perform local DNS resolution when possible using a local DNS resolution database

Provide support for cookies when needed using a cookie database

Eliminate duplicate URLs with Dupserver 224

Fig. 3A

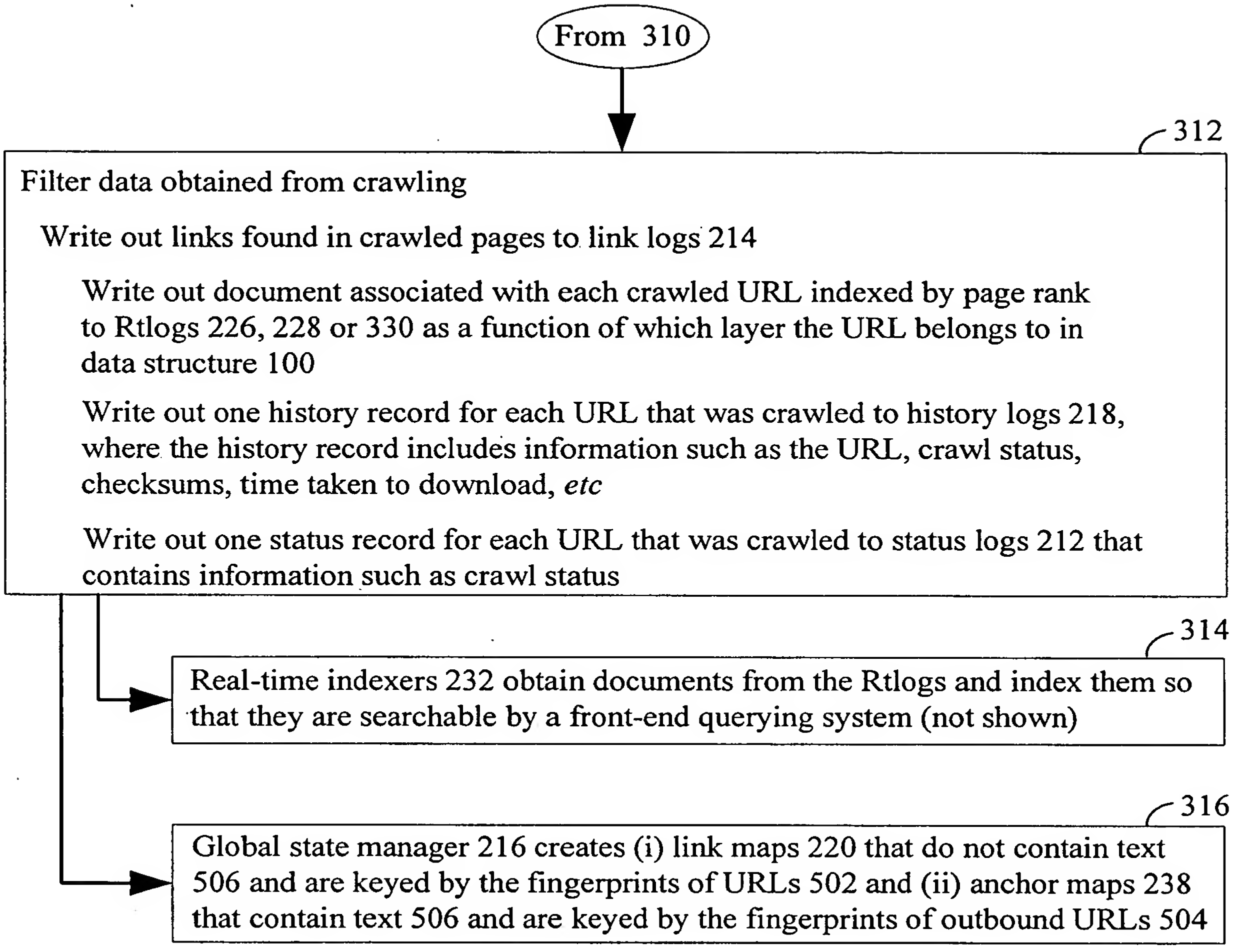

From 310
312
Filter data obtained from crawling
Write out links found in crawled pages to link logs 214
Write out document associated with each crawled URL indexed by page rank to Rtlogs 226, 228 or 330 as a function of which layer the URL belongs to in data structure 100
Write out one history record for each URL that was crawled to history logs 218, where the history record includes information such as the URL, crawl status, checksums, time taken to download, *etc*
Write out one status record for each URL that was crawled to status logs 212 that contains information such as crawl status
314
Real-time indexers 232 obtain documents from the Rtlogs and index them so that they are searchable by a front-end querying system (not shown)
316
Global state manager 216 creates (i) link maps 220 that do not contain text 506 and are keyed by the fingerprints of URLs 502 and (ii) anchor maps 238 that contain text 506 and are keyed by the fingerprints of outbound URLs 504

Fig. 3B

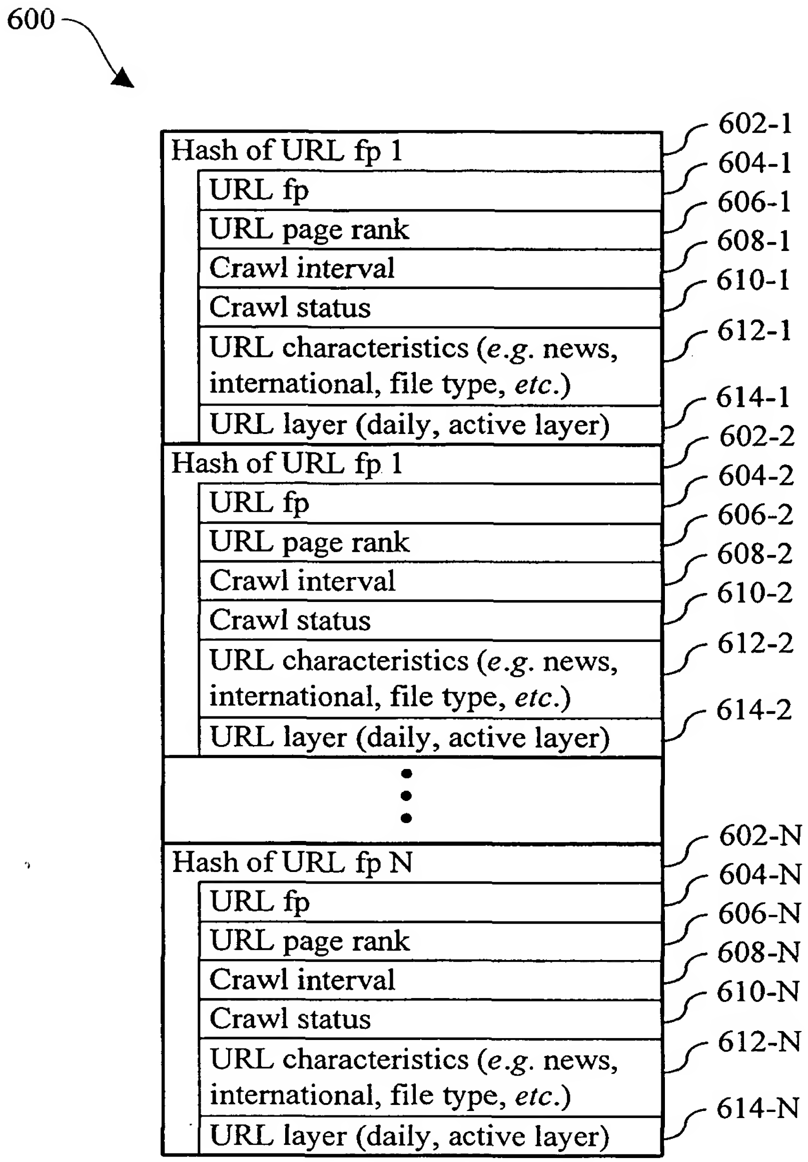

600
Hash of URL fp 1
URL fp
URL page rank
Crawl interval
Crawl status
URL characteristics (*e.g.* news, international, file type, *etc.*)
URL layer (daily, active layer)
602-1
604-1
606-1
608-1
610-1
612-1
614-1
Hash of URL fp 1
URL fp
URL page rank
Crawl interval
Crawl status
URL characteristics (*e.g.* news, international, file type, *etc.*)
URL layer (daily, active layer)
602-2
604-2
606-2
608-2
610-2
612-2
614-2
Hash of URL fp N
URL fp
URL page rank
Crawl interval
Crawl status
URL characteristics (*e.g.* news, international, file type, *etc.*)
URL layer (daily, active layer)
602-N
604-N
606-N
608-N
610-N
612-N
614-N

Fig. 4

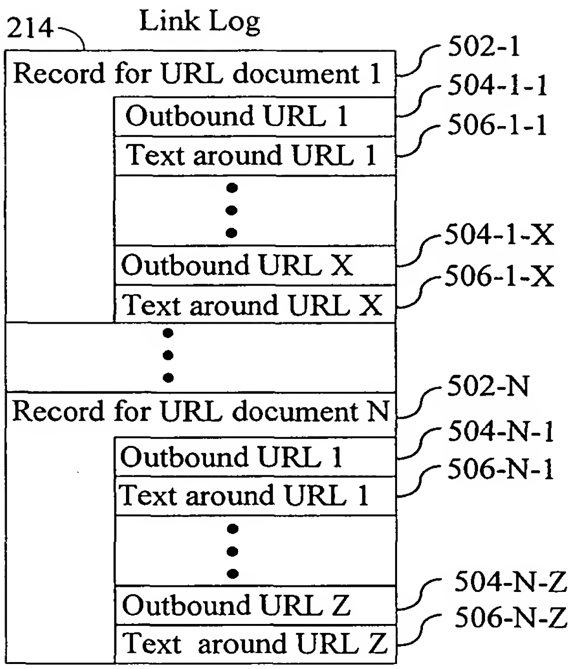


Fig. 5A

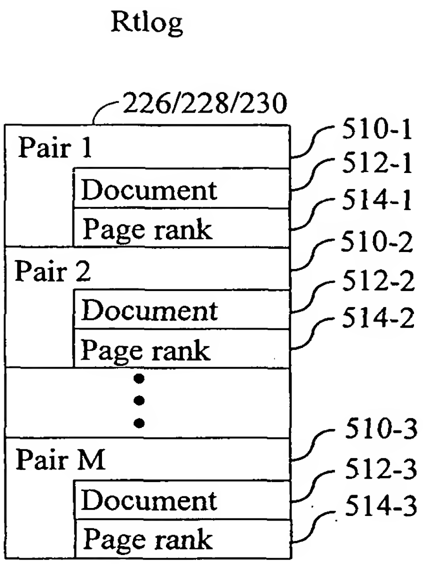


Fig. 5B

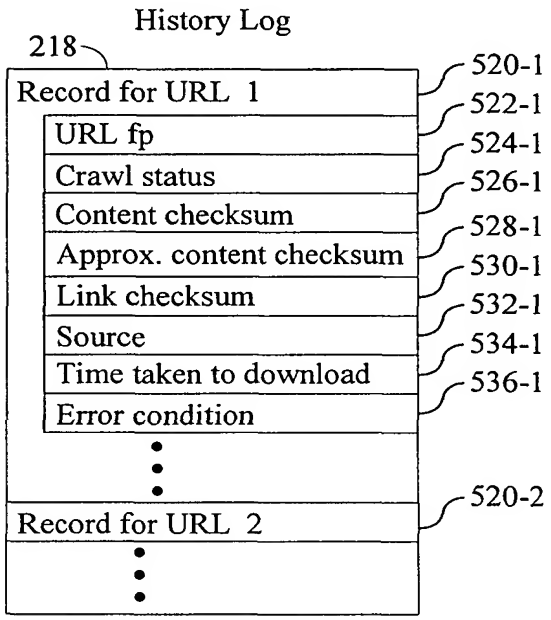


Fig. 5C

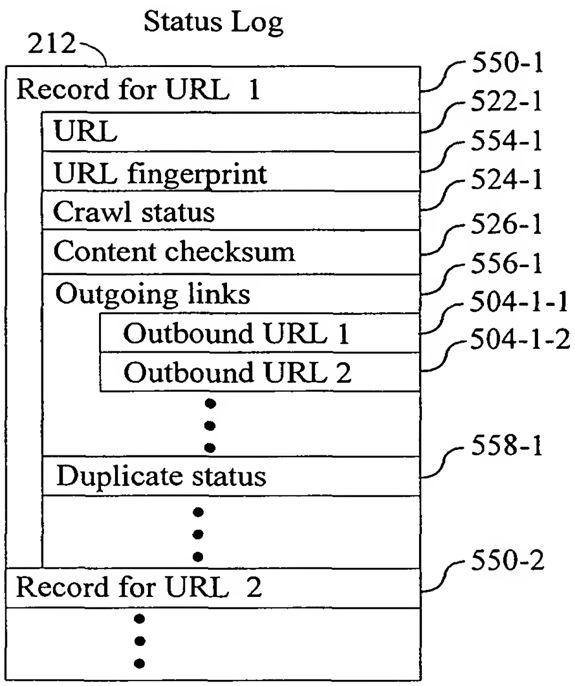


Fig. 5D

Fig. 5

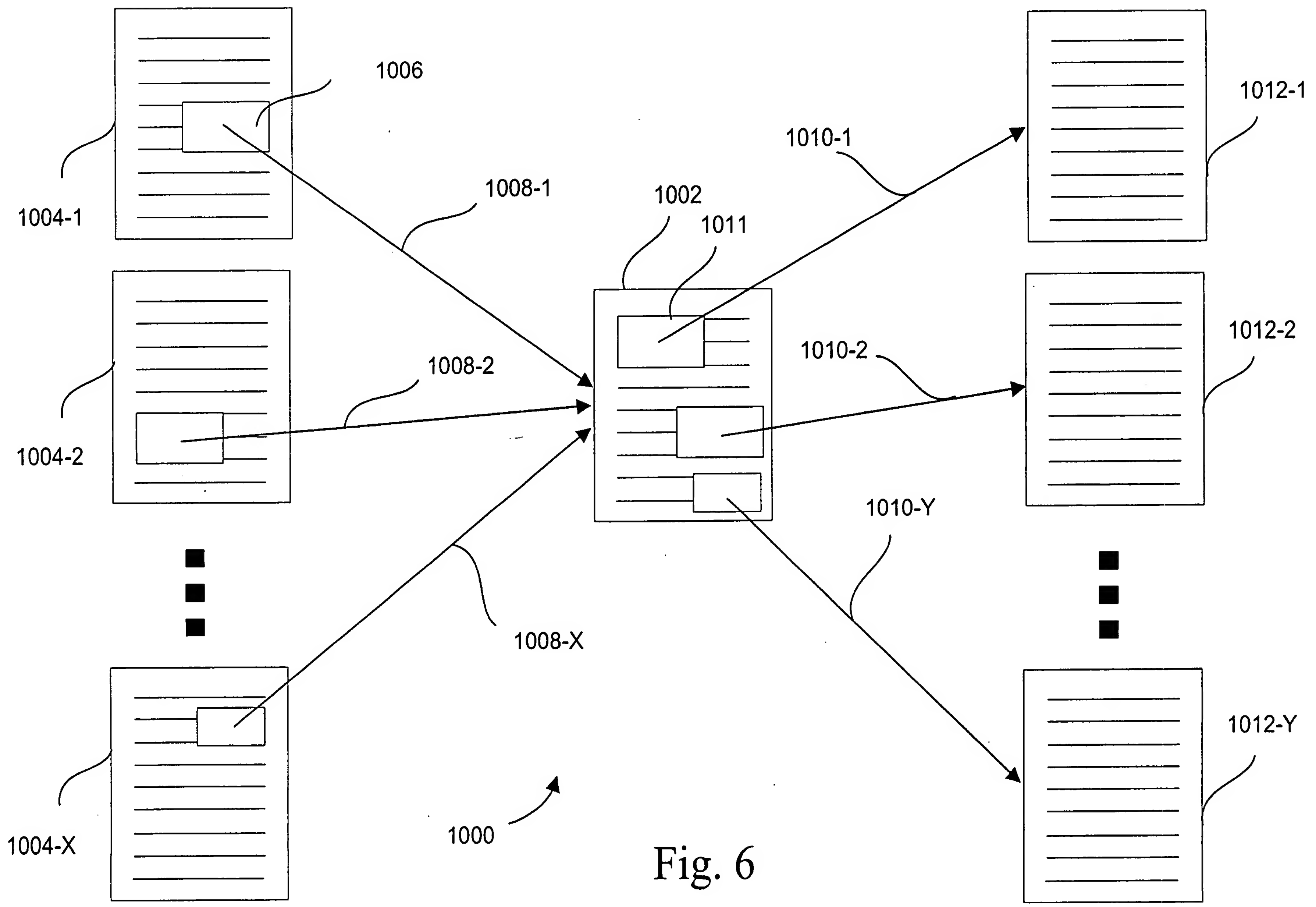

1006
1004-1
1008-1
1002
1011
1010-1
1012-1
1008-2
1004-2
1010-2
1012-2
1010-Y
1008-X
1004-X
1012-Y
1000

Fig. 6

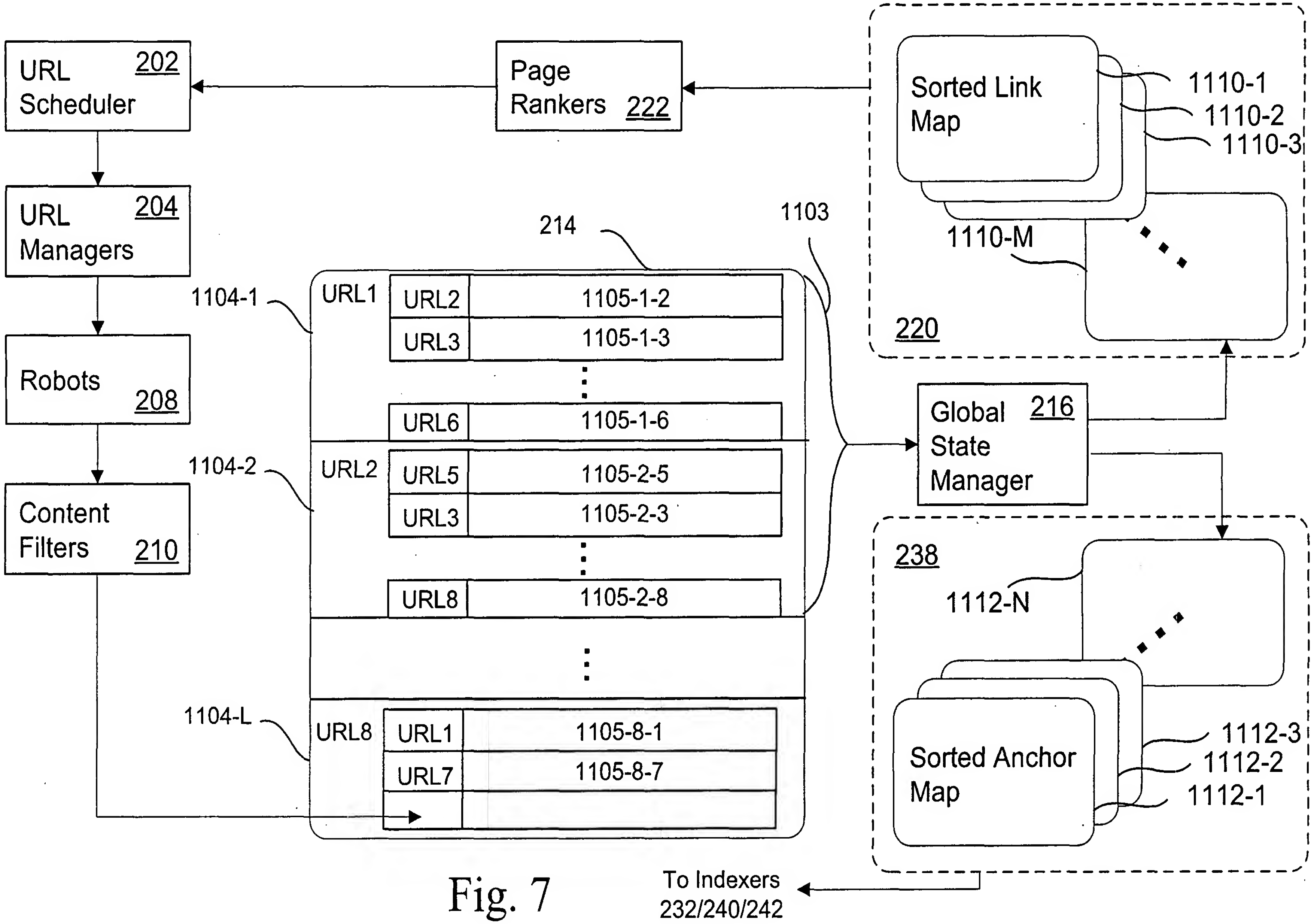

URL Scheduler 202
URL Managers 204
Robots 208
Content Filters 210
Page Rankers 222
Sorted Link Map
1110-1
1110-2
1110-3
1110-M
220
214
1103
1104-1
URL1
URL2
1105-1-2
URL3
1105-1-3
URL6
1105-1-6
1104-2
URL2
URL5
1105-2-5
URL3
1105-2-3
URL8
1105-2-8
1104-L
URL8
URL1
1105-8-1
URL7
1105-8-7
Global State Manager 216
238
1112-N
Sorted Anchor Map
1112-3
1112-2
1112-1
To Indexers 232/240/242

Fig. 7

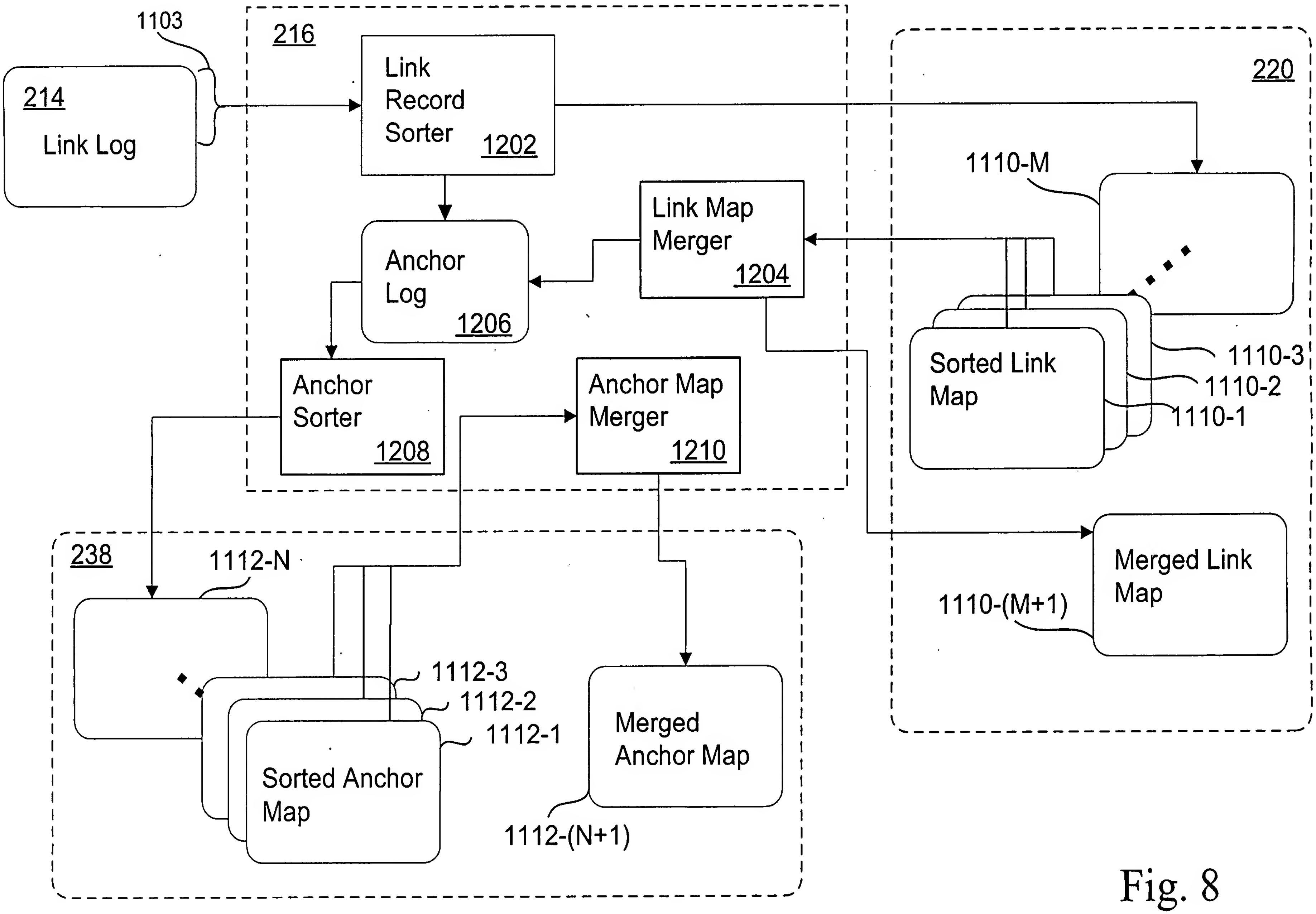
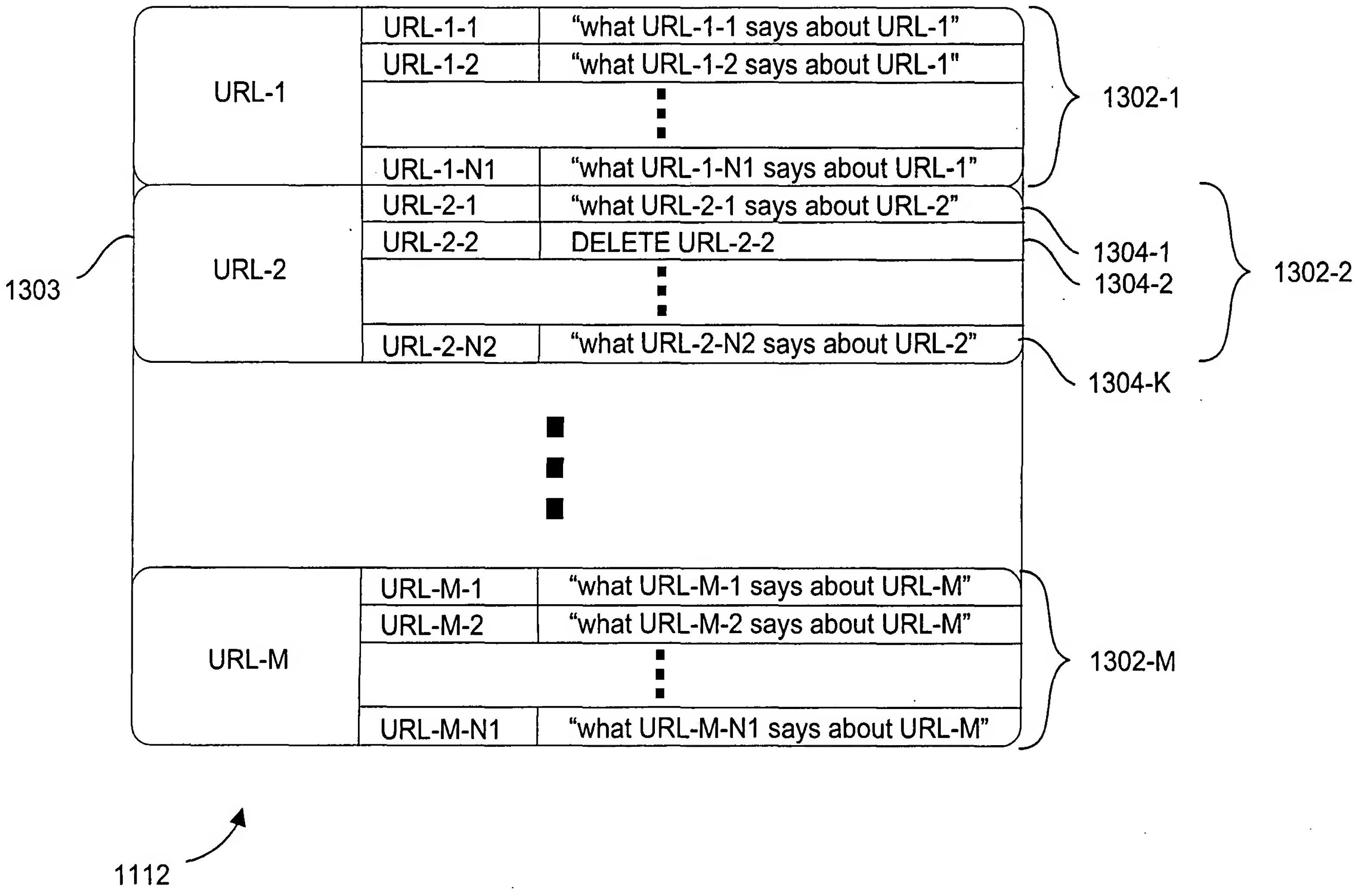

1103
214
Link Log
216
Link Record Sorter
1202
Anchor Log
1206
Link Map Merger
1204
Anchor Sorter
1208
Anchor Map Merger
1210
220
1110-M
Sorted Link Map
1110-3
1110-2
1110-1
Merged Link Map
1110-(M+1)
238
1112-N
1112-3
1112-2
1112-1
Sorted Anchor Map
Merged Anchor Map
1112-(N+1)

Fig. 8



URL-1	URL-1-1	“what URL-1-1 says about URL-1”
	URL-1-2	“what URL-1-2 says about URL-1"
	⋮	
	URL-1-N1	“what URL-1-N1 says about URL-1”
URL-2	URL-2-1	“what URL-2-1 says about URL-2”
	URL-2-2	DELETE URL-2-2
	⋮	
	URL-2-N2	“what URL-2-N2 says about URL-2”
⋮		
URL-M	URL-M-1	“what URL-M-1 says about URL-M”
	URL-M-2	“what URL-M-2 says about URL-M”
	⋮	
	URL-M-N1	“what URL-M-N1 says about URL-M”

Fig. 9

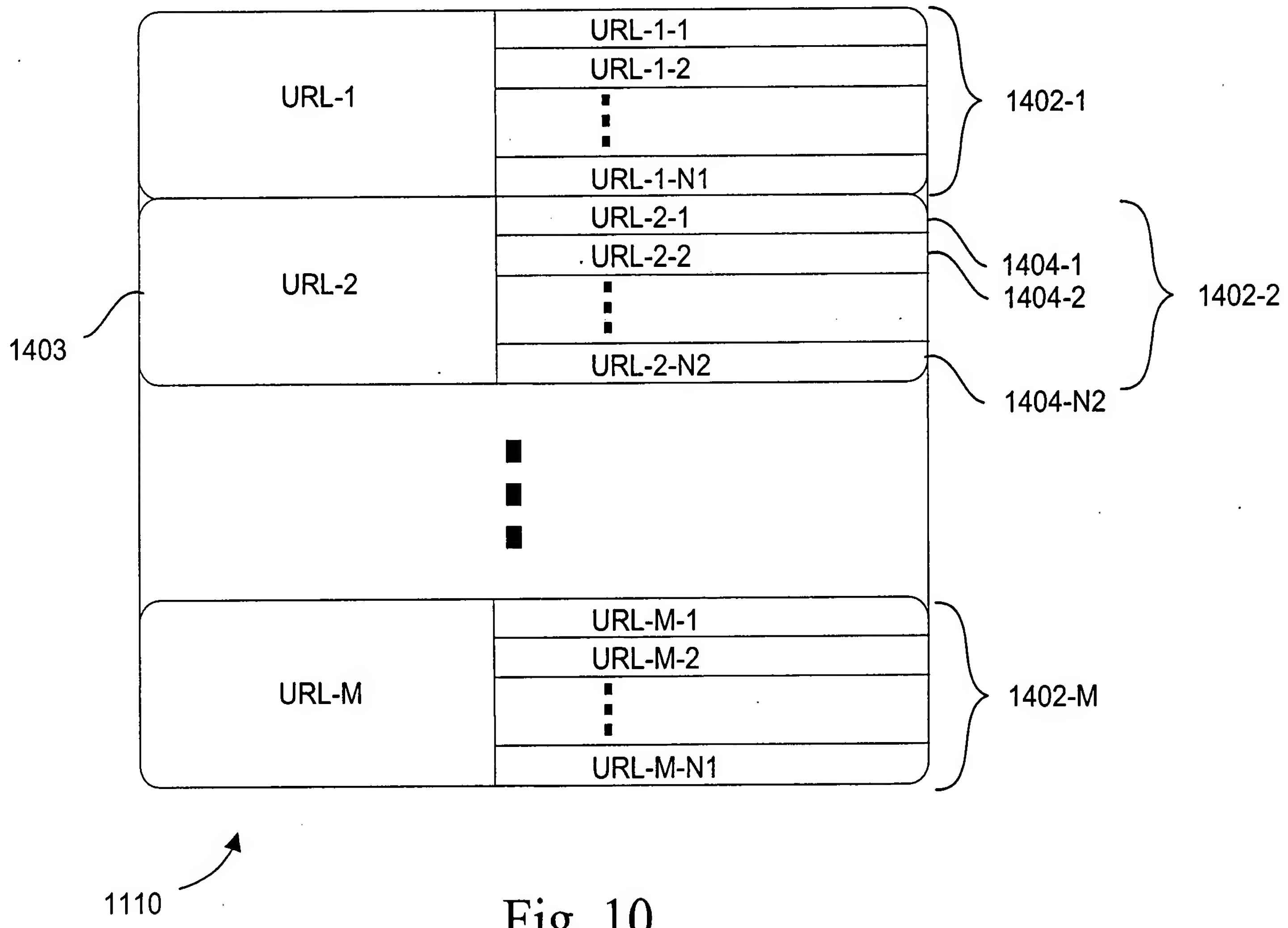

URL-1
URL-1-1
URL-1-2
URL-1-N1
1402-1
URL-2
URL-2-1
URL-2-2
URL-2-N2
1403
1404-1
1404-2
1402-2
1404-N2
URL-M
URL-M-1
URL-M-2
URL-M-N1
1402-M
1110

Fig. 10

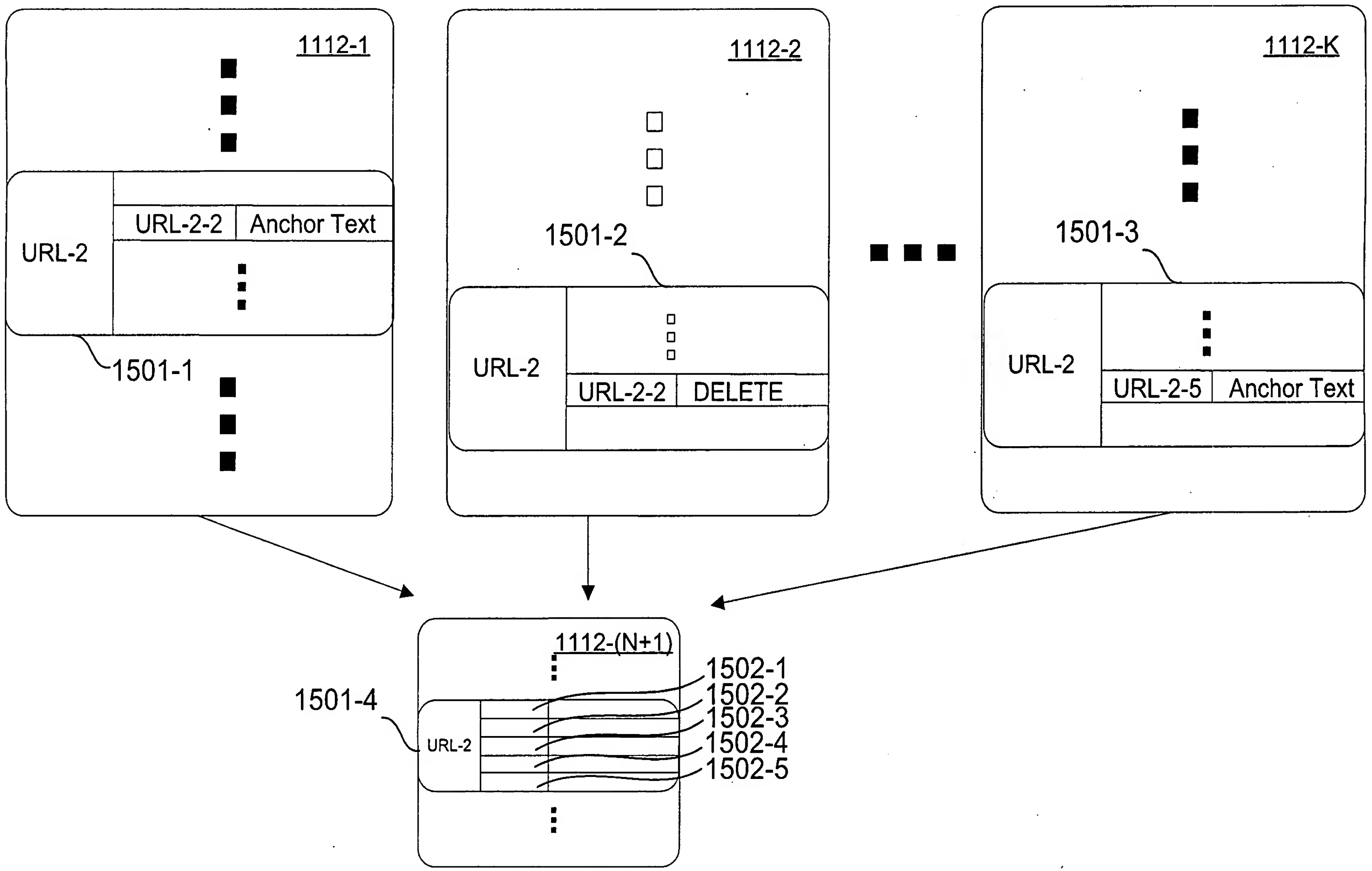

1112-1
URL-2
URL-2-2
Anchor Text
1501-1
1112-2
1501-2
URL-2
URL-2-2
DELETE
1112-K
1501-3
URL-2
URL-2-5
Anchor Text
1112-(N+1)
1501-4
URL-2
1502-1
1502-2
1502-3
1502-4
1502-5

Fig. 11

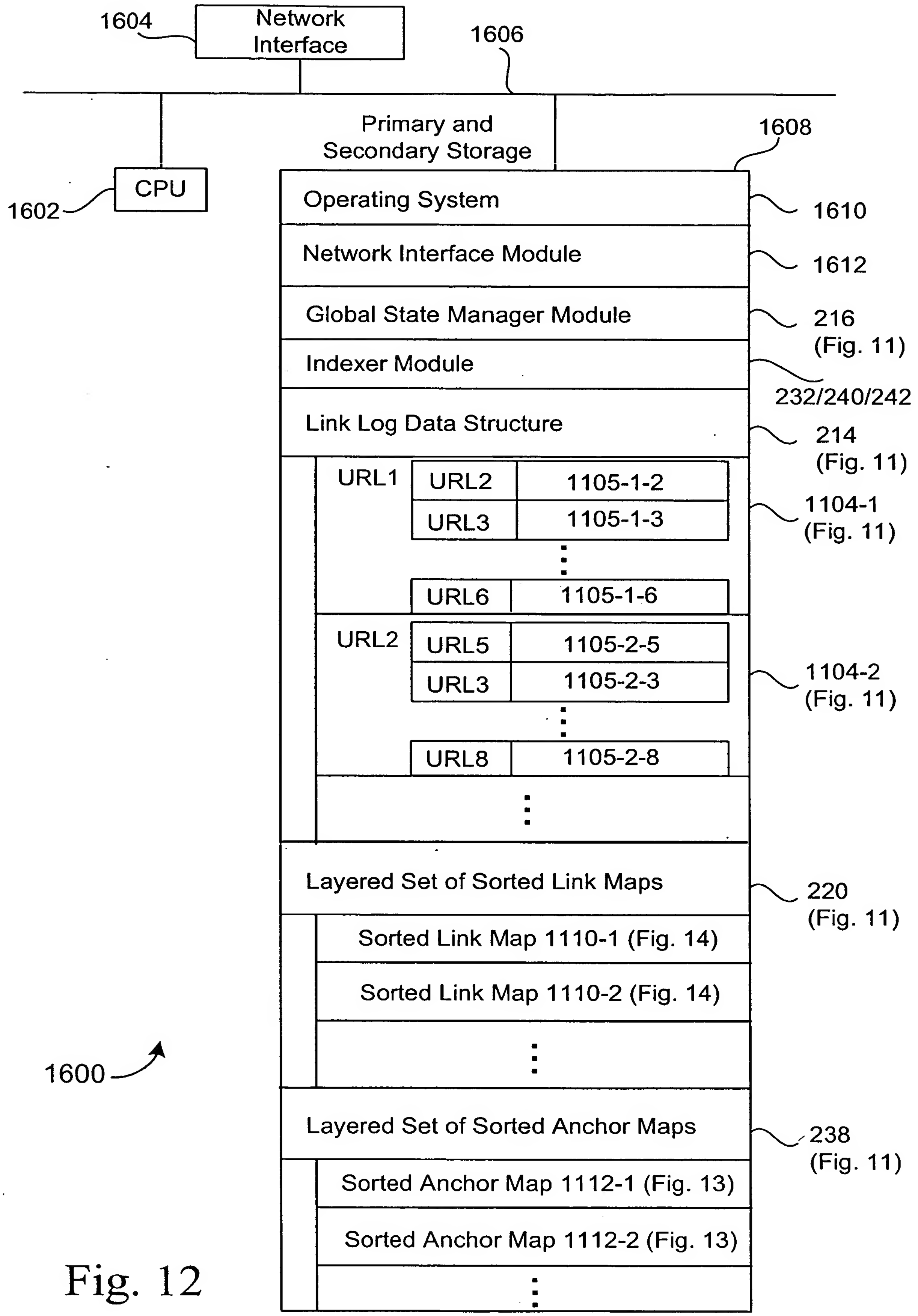

1604
Network Interface
1606
Primary and Secondary Storage
1608
CPU
1602
Operating System
1610
Network Interface Module
1612
Global State Manager Module
216 (Fig. 11)
Indexer Module
232/240/242
Link Log Data Structure
214 (Fig. 11)
URL1
URL2
1105-1-2
URL3
1105-1-3
URL6
1105-1-6
1104-1 (Fig. 11)
URL2
URL5
1105-2-5
URL3
1105-2-3
URL8
1105-2-8
1104-2 (Fig. 11)
Layered Set of Sorted Link Maps
220 (Fig. 11)
Sorted Link Map 1110-1 (Fig. 14)
Sorted Link Map 1110-2 (Fig. 14)
1600
Layered Set of Sorted Anchor Maps
238 (Fig. 11)
Sorted Anchor Map 1112-1 (Fig. 13)
Sorted Anchor Map 1112-2 (Fig. 13)

Fig. 12